Status Report

The Feasibility of Publishing County-level Estimates of the Number of Women Eligible for the CDC's NBCCEDP

U.S. Census Bureau

Demographic Directorate

Small Area Health Insurance Estimates Team

May 2008, revised October 2008

1. Introduction

This is a status report on the U.S. Census Bureau's Small Area Health Insurance Estimates (SAHIE) program's research on estimating the number of women eligible to participate in the Centers for Disease Control and Prevention's (CDC) National Breast and Cervical Cancer Early Detection Program (NBCCEDP). This report provides an assessment of which county-level estimates are feasible to produce and publish on the Census Bureau's web site at the end of Phase Four of this study (Fall 2008).

Phase One dealt primarily with data acquisition and model development for the uninsured. Phase Two assessed the feasibility of producing estimates of the number of low-income, uninsured women by age, race, and ethnicity. Phase Three developed a state-level model. Early in Phase Four (July 2007), the SAHIE program published state-level estimates of the number of uninsured, low-income women by age, race, and ethnicity. Low-income is defined as ≤ 200% or ≤ 250% of the federal poverty threshold. We published estimates for both categores for each state. The age categories are 18-64 years, 40-64 years, and 50-64 years. The race and ethnicity categories are All Races, non-Hispanic White, non-Hispanic Black, and Hispanic.

Published estimates and documentation of our methodology are available on the SAHIE program's web site, <http://www.census.gov/hhes/www/sahie/>.

After releasing the state-level estimates in July 2007, the SAHIE team has developed and refined the county model. The knowledge gained from the state model proved invaluable for formulating the county model. Additionally, lessons are being learned from the county model that will improve the state model for future releases.

In 2006, the SAHIE team provided an initial assessment of feasible estimates.[1] We have now affirmed our initial opinion that a county-level model could neither accommodate race and ethnicity categories nor jointly model estimates of both the 200% and 250% income-to-poverty ratios (IPRs) within the same model. So, it is currently feasible to produce estimates of either ≤ 200% IPR or ≤ 250% IPR for counties, but not both. This report will address all states using the ≤ 200% IPR; if the model is successful at the 200% IPR, we expect that the estimates will be relatively more precise at the 250% IPR for those states needing this IPR.

In summary, the SAHIE team has determined that it is feasible to produce and publish county-level estimates of the number of uninsured, low-income women by age. For this study, low-income is defined as either ≤ 200% or ≤ 250% of the federal poverty threshold; we will provide the relevant poverty threshold, either 200% or 250%, for each state. The feasible age categories are 18-64 years and 40-64 years. There is still uncertainty as to whether the precision of the estimates for women aged 50-64 is adequate. This will be addressed in the following sections.

Section 2 discusses the quality standards related to publishing estimates; Section 3 presents a summary of the findings evaluating whether the precision of the estimates meets the Census Bureau's and the CDC's standards; Section 4 discusses the results in detail; and Section 5 provides our conclusions.

2. Quality Standards

To determine if the quality of an estimate is adequate for publication, the SAHIE team evaluated the precision of the estimate. Precision can be measured in terms of variances, standard errors, half-lengths of confidence intervals, or the coefficients of variation (CVs). We assess the feasibility of producing publication quality estimates of uninsured, low-income women in each of three age groups by evaluating whether the precision of the estimates meets Census Bureau and CDC standards on estimated CVs.

In 2007, the Census Bureau published quality requirements for releasing data products.[2] These quality standards apply "to all releases of data by the Census Bureau using the results of surveys or censuses" and address nonsampling error and sampling error. While these standards do not specifically address model-based estimates, we can look to them for guidance. The standards are in transition, and the Census Bureau is actively working on a standard for modeled estimates.

[1] Brett O'Hara, Joanna Turner, Mark Bauder, Steven Riesz, and David Waddington (2006), "Initial Assessment of Small Area Estimation of the Number of Eligible Women for the CDC's NBCCEDP," available at <http://www.census.gov/hhes/www/sahie/pubs/feasibilityph2.pdf>.

[2] U.S. Census Bureau (2007), "Quality Requirements for Releasing Data Products," available at <http://www.census.gov/quality/S20-0_v1.0_Data_Release.pdf>.

For the sampling error standard, key statistics need to be identified. For example, estimates of the number of low-income, uninsured women aged 18-64, 40-64, and 50-64 are key statistics for this project. To satisfy the Public Data Release Criteria for a survey, a majority of the released key statistics in a product should have an estimated CV of 30 percent or lower. Since this standard need not apply to every released estimate, the Census Bureau may publish estimates for which the estimated CV is higher than 30 percent as long as the majority of the CVs for the key statistics are 30 percent or lower. The distribution and medians of the estimated CVs of key statistics for the counties can be examined to assess whether this requirement has been met.

The CDC has not informed the SAHIE team members what level of precision is adequate to serve their purpose. We do know that the CDC does not publish an individual survey estimate if the estimated CV (i.e., the relative standard error) is higher than 30 percent.[3] Although these standards do not directly apply to model-based estimates, they serve as an upper bound for the CV rules for survey estimates.

The SAHIE program models health insurance coverage as measured by the Annual Social and Economic Supplement of the Current Population Survey (CPS ASEC). There are a few caveats about comparing the county estimates in this report (based on 2001-2003 CPS ASEC data) with future estimates that will be up-to-date in Fall 2008 (based on 2005-2007 CPS ASEC data).[4] The largest difference is that this report shows health insurance estimates that are based on unrevised CPS ASEC data.[5] Estimates based on revised CPS ASEC data show fewer uninsured people than estimates based on unrevised data. Using unrevised data may raise the CV on the uninsured population. Also, when producing 2005 as opposed to 2001 model-based estimates, we will have time lagged data for Census 2000, Medicaid, and probably other variables that could be used in our model to predict the uninsured population. The larger the time gap between the CPS ASEC and predictive variables, the variables are less predictive and the variances (and CVs) become higher.

In addition to the caveats just described, there are a few cautions on how the final production estimates for counties will differ from the estimates in this report. First, the models and estimates in this report are not final. Some evidence of lack of fit of the county model still needs to be resolved. Second, the estimates have not been controlled to the state estimates; the state controls help to minimize possible bias in the county point estimates because the state estimates are based on more reliable direct estimates. Third, the correlation of the direct estimates between counties within a state has not yet been incorporated into the model because methods for doing this are still being developed. Fourth, the variances for the direct estimates have been implicitly estimated in the model; we are developing direct estimates of these variances. Fifth, the correlation of the direct

[3] Klein RJ, Proctor SE, Boudreault MA, Turczyn KM, “Healthy People 2010 criteria for data suppression,” Statistical Notes, no 24. Hyattsville, Maryland: National Center for Health Statistics, June 2002; page 9.

[4] We use the 2001-2003 CPS ASECs to produce estimates for 2001 and the 2005-2007 CPS ASECs to produce estimates for 2005.

[5] The CPS ASEC data used for our 2001 model-based estimates were tabulated prior to the revision. For more details on the revision, see <Hhttp://www.census.gov/hhes/www/hlthins/usernote/schedule.html>.

estimates between different age, sex, and IPR groups has not been incorporated into the model yet; work has begun on estimating these correlations. These issues may raise the variances in the final production estimates, but it is also possible that further improvements in the model might reduce these variances. On net, there is no specific way to anticipate all of these effects on the variances (and CVs).

3. Summary of Findings

We have confirmed that the SAHIE program can produce estimates for uninsured, low-income ($\leq$ 200% IPR) women aged 18-64 and 40-64 and that these estimates meet the Census Bureau's standard that at least half of the estimated CVs be 30 percent or less. In addition, under the CDC's reliability criteria for data suppression, virtually all of the estimates for women aged 18-64 could be published and about 84 percent of the estimates for women aged 40-64 could be published. The level of precision for women aged 50-64 is markedly less than that for women aged 40-64 and may not be adequate for the CDC's purpose. Even though we could produce estimates for women aged 50-64, the estimates would be *at best* marginally publishable under the Census Bureau's standards and less than half of the estimates would meet the CDC's criterion that estimated CVs must be 30 percent or less.

For each age group, 40-64 years or 50-64 years, the CVs have a similar distribution (median and dispersion) for the majority of the estimates, and the median is smaller for the larger estimates, the larger proportions, and the counties with larger populations. The relationship of the CV to the proportion of low-income women who are uninsured is tighter than that of the CV to the proportion of women who are both low-income and uninsured (i.e., the eligible population for NBCCEDP services). The half-lengths of confidence intervals are nearly proportional to the estimates except for the largest estimates.

4. Results

In this section, we present results from a preliminary version of our model. We analyze the CVs of the number of uninsured, low-income women for the key age groups. The CVs in this report are the posterior standard errors of the numbers of uninsured, low-income women divided by their posterior means.[6] As described in Section 2, the CVs we report here may differ from those we will report in Fall 2008. Section 4.1 analyzes the distributions of the CVs for the age groups 18-64, 40-64, and 50-64. Section 4.2 analyzes the CVs in relationship to different measurements of uninsured, low-income women and

[6] We have used a Bayesian modeling approach. This approach determines the conditional distributions of the numbers uninsured after data have been observed, hence posterior distributions. The means and standard errors from these posterior distributions are used to calculate the CVs.

against their county's population. Section 4.3 presents an analysis of the expected half-confidence widths.

We do not know at this time how to estimate the sampling variances of the estimated CVs in this report nor the correlations between these CVs induced by the modeling. Consequently, any comparisons made in this report have not been tested for statistical significance.

4.1 Distributions of Coefficients of Variation

Table 1 summarizes the distributions of the estimated CVs for the estimates for the three age groups. The medians are 0.221 and 0.249 for women 18-64 and 40-64, respectively, which are well below 0.30. For women 50-64, the median is barely above 0.30.

Table 1. Estimated coefficients of variation for county-level estimates of the number of uninsured, low-income women for key age groups.

		Percentiles					
Age	Min	10	25	Median	75	90	Max
18-64	0.051	0.157	0.186	0.221	0.256	0.281	0.406
40-64	0.076	0.179	0.210	0.249	0.285	0.314	0.477
50-64	0.108	0.217	0.257	0.301	0.344	0.381	0.545

Table 2 shows the proportion of estimated CVs that are below selected thresholds. For women aged 18-64, about one-third of the CVs are below 0.20, and only a small proportion (4%) have CVs above 0.30 percent. For women aged 40-64, about one-fifth are below 0.20, about 84% are below 0.30, and nearly all are below 0.40. The CVs for women aged 50-64 are much larger. As noted above, less than half of the CVs are less than 0.30, few are less than 0.20, and more than 20% are above 0.35. Figure 1 graphically shows the distributions of CVs for ages 40-64 and 50-64.

Table 2. Proportions of estimated CVs below selected thresholds for county-level estimates of the number of uninsured, low-income women for key age groups.

	Proportion of Estimated CVs					
Age	≤ 0.15	≤ 0.20	≤ 0.25	≤ 0.30	≤ 0.35	≤ 0.40
18-64	0.073	0.334	0.709	0.962	0.998	1.000
40-64	0.029	0.198	0.504	0.836	0.974	0.998
50-64	0.006	0.057	0.218	0.494	0.785	0.943

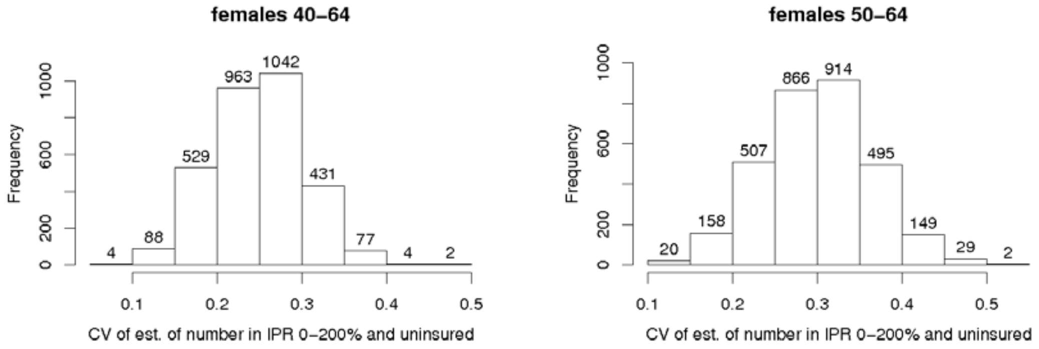


Figure 1. Histogram of estimated CVs of estimates of the number of low-income, uninsured women.

4.2 Analysis of Estimated CVs for Women Aged 40-64 and 50-64

Because almost all of the estimates for women 18-64 years old satisfy the 30 percent CV threshold, the remainder of this analysis will focus on estimates for uninsured, low-income women aged 40-64 and 50-64. The scatter plots include loess curves that show the central trends in each of the plots. Loess is a nonparametric method to fit flexible curves that make no assumptions on their shape. It can show how the central tendencies of the plotted data change. If the loess curve is nearly horizontal in part of the plot then there is little or no trend. An increasing or decreasing curve shows a positive or negative trend in that part of the plot. A plot with few points would not represent a reliable trend.

The plots in Figure 2 indicate downward trends. This means that the standard error is decreasing faster than its estimate, so that large estimates of the number of uninsured, low-income women have relatively smaller standard errors. Figure 3 contains box plots of the estimated CVs for approximate quintiles of the number of uninsured, low-income women.[7] For the first four quintiles, estimates of 800 or fewer women aged 40-64, the medians are about the same. The median for the fifth quintile is somewhat smaller because it contains counties with a non-negligible sample that would dominate the decrease. About 60 percent of counties have no sample in the CPS ASEC and most of the rest have small samples, so that the model instead of the direct survey estimates is dominant. Counties that have little or no sample would mostly be small and would be expected to have the smallest numbers of uninsured, low-income women. Similarly for women aged 50-64, the first four quintiles, estimates of 400 or fewer, have similar medians, and the fifth quintile has the smallest median. For most of the quintiles, these women have a greater dispersion (larger interquartile range) than women aged 40-64.

[7] The line in the middle of the box plot is the median, and the distance between the top and bottom of the box is the interquartile range, the difference between the 75th percentile and the 25th percentile, which is a measure of dispersion.

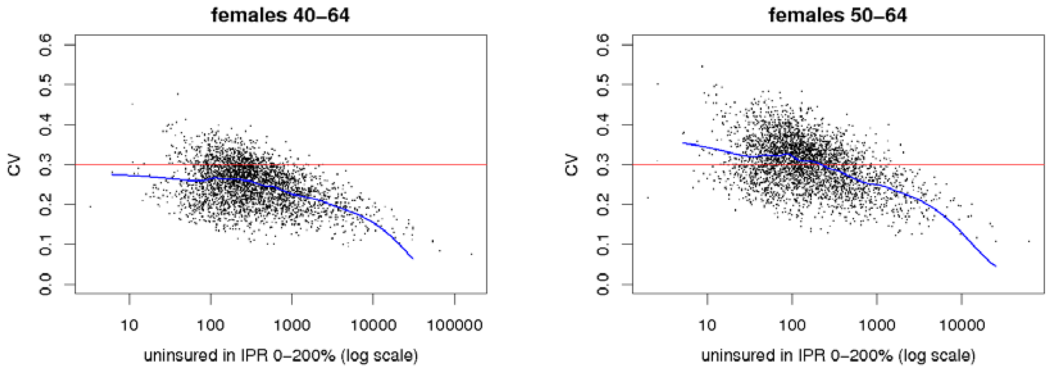


Figure 2. Estimated CVs of estimates by county of the number of uninsured, low-income women plotted against the number of uninsured, low-income women.

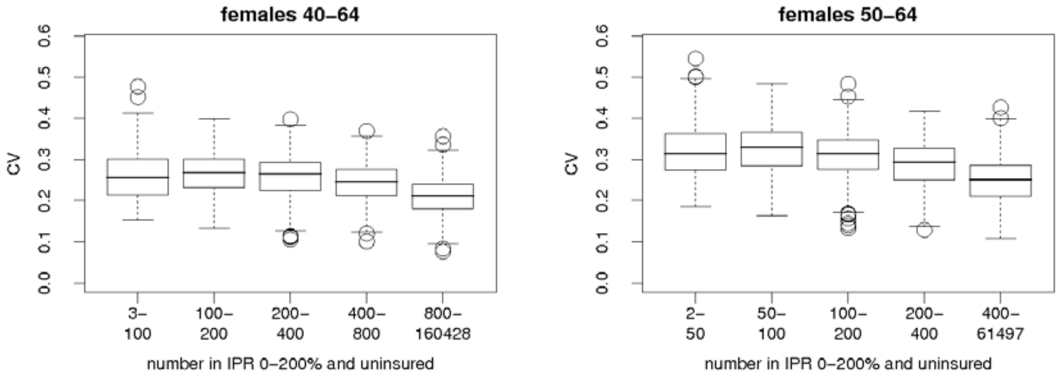


Figure 3. Boxplots of estimated CVs of estimates of number of uninsured, low-income women plotted against the number of uninsured, low-income women: groups are approximately quintiles.

Figure 4 plots estimated CVs against the logit of the proportion of women who are both low-income and uninsured, the screening population. The logit is the log of the proportion divided by one minus the proportion, i.e., $\log\left(p/(1-p)\right)$ where p is the proportion. Even though we are plotting the CVs of the number of uninsured, low-income women against the proportions, there is a downward trend showing an inverse relationship to the proportions. Figure 5 shows box plots of the estimated CVs for approximate quintiles of the proportion of women who are both low-income and uninsured. The trend shows an apparent steady decline. There is a small trend in the CVs for the first three quintiles and a decrease is evident in the fifth quintile. The fifth quintile contains the largest proportion of low-income, uninsured women. Keep in mind that we cannot measure statistical significance between these apparent trends.

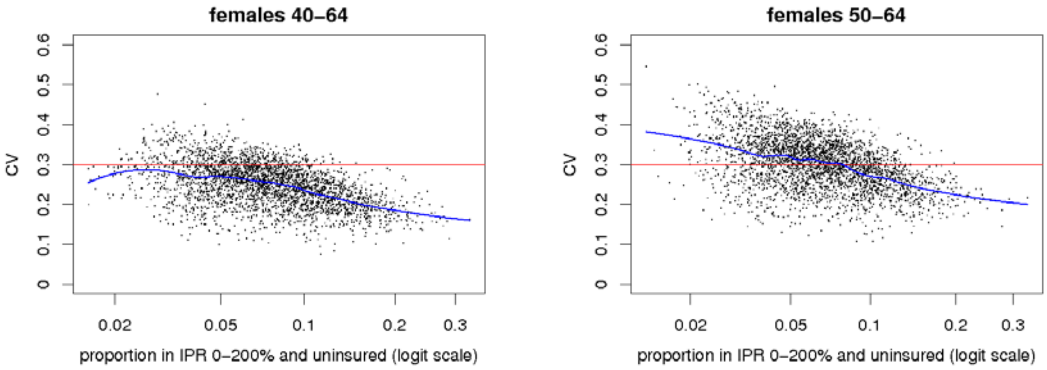


Figure 4. Estimated CVs of estimates by county of the number of uninsured, low-income women plotted against the proportion of women who are both low-income and uninsured.

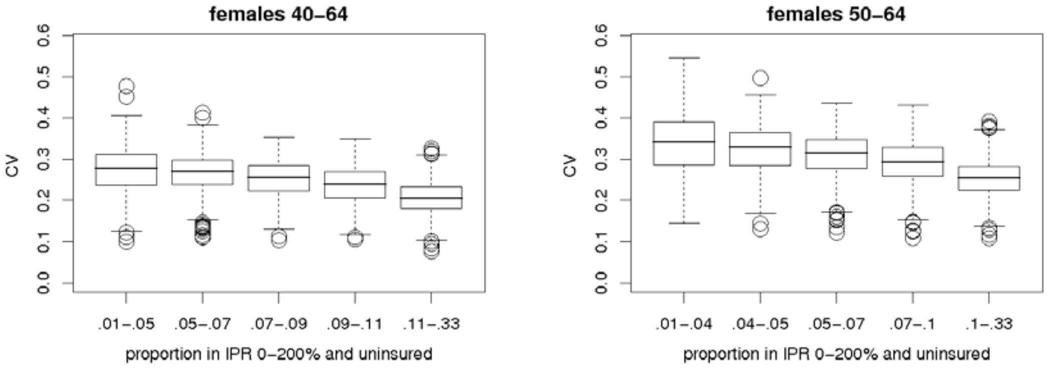


Figure 5. Boxplots of estimated CVs of estimates by county of the number of uninsured, low-income women plotted against the proportion of women who are both low-income and uninsured.

Figure 6 plots estimated CVs against a different proportion than Figure 4. It plots the CVs against the logit of the proportion of low-income women that are uninsured. It shows a tighter relationship between the proportions and the CVs and clearly shows that the CVs decrease with these increasing proportions. The boxplots in Figure 7 confirm the downward trend and the much tighter relationship throughout the full range of the proportions.

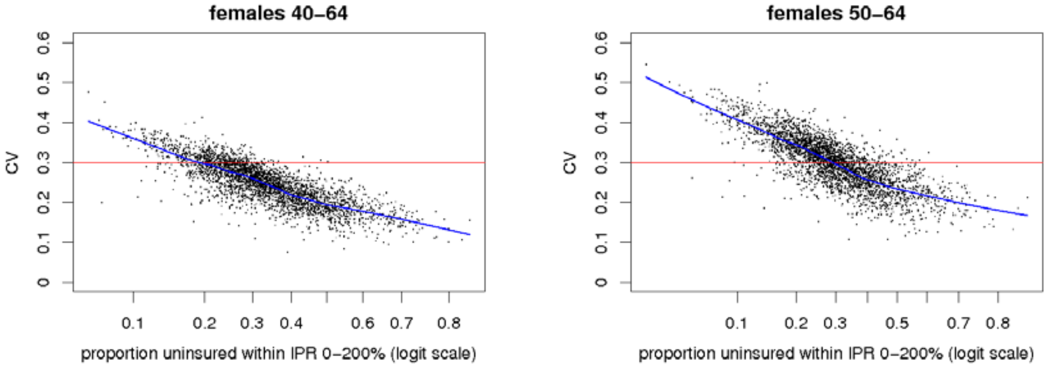


Figure 6. Estimated CVs of estimates by county of the number of uninsured, low-income women plotted against the proportion of low-income women that are uninsured.

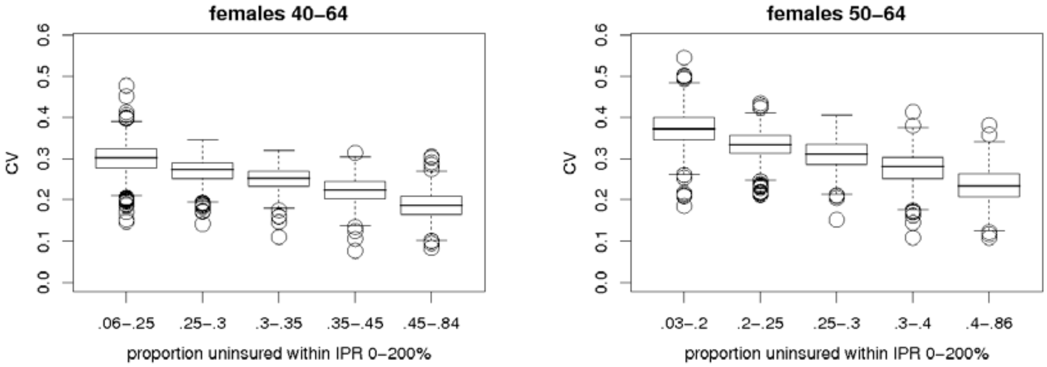


Figure 7. Boxplots of estimated CVs of estimates of number of uninsured, low-income women plotted against proportion of low-income women that are uninsured. Groups are approximately quintiles.

Figure 8 shows large dispersion in the estimated CVs by county with no clear relationship between the CVs and the county population for most of the counties. The largest counties tend to have smaller CVs. This might be due to non-negligible sample sizes or larger proportions of uninsured, low-income women.

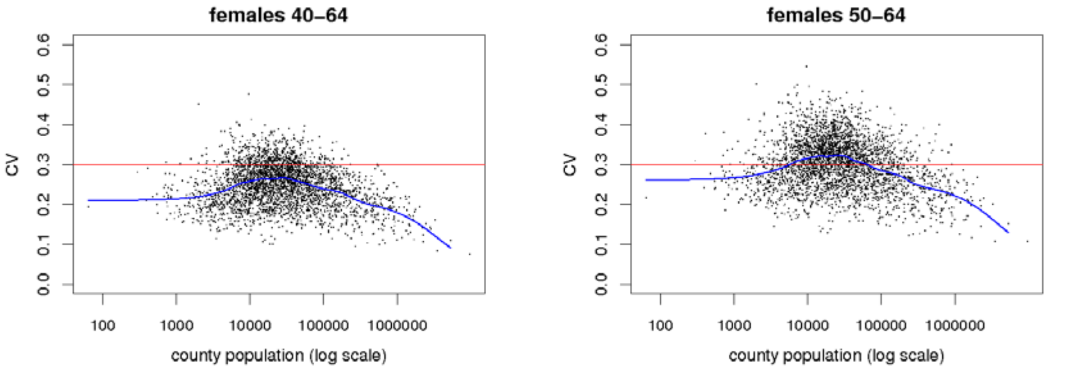
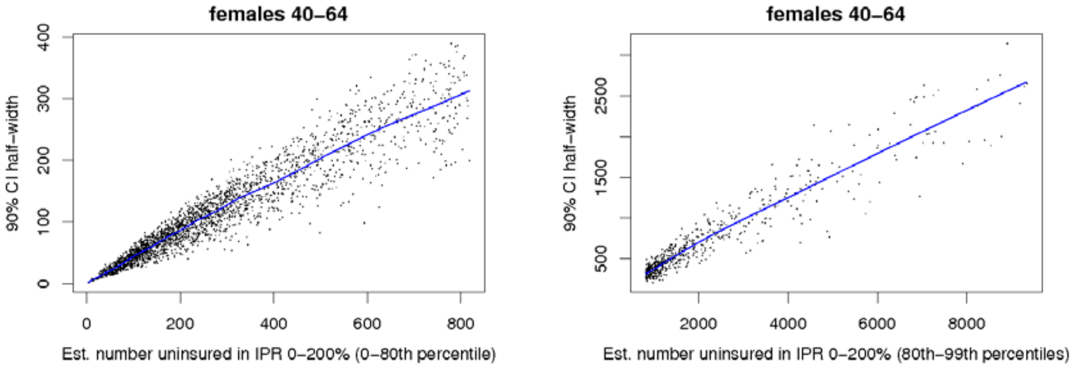


Figure 8. Estimated CVs of estimates by county of the number uninsured, low-income women plotted against county population.

4.3 Analysis of Half Confidence Widths for Women Aged 40-64 and 50-64

Figure 9 has plots of the half confidence widths (1.645 × estimated standard error) for 90% confidence intervals for the estimate of the number of uninsured, low-income women plotted against the estimate. The top plots are for women aged 40-64 and the bottom plots are for women aged 50-64. The left-side plots include counties with the smallest estimates; those in the first four quintiles (0 – 80th percentile) and the right-side plots are for the largest estimates; those in the top quintile exclude the largest 1%. In the left-side plots, the loess curves, which give a sense of the expected half-confidence width, show a slight negative curvature, indicating that the confidence width increases are less than proportional to the estimates. This would be expected based on the analysis of Figure 2 and Figure 3, where the estimated CVs tend to be smaller for the larger estimates. The right-side plots for the largest estimates do not exhibit a departure from proportionality.



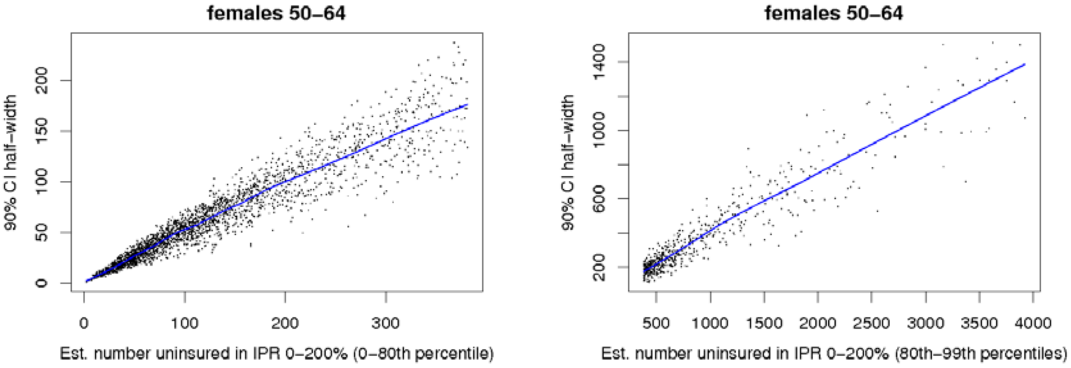


Figure 9. Length of 90% half-width confidence interval for number of uninsured, low-income women.

5. Conclusion

At the end of Phase Four (Fall 2008), we will be able to publish on the Census Bureau's web site county estimates for uninsured, low-income (≤ 200% or ≤ 250% IPR) women aged 18-64 and 40-64. [8]

Our current research into county estimates for uninsured, low-income women aged 50-64 indicates that they (marginally) lack the precision required for publication on the Census Bureau's web site. As discussed in Section 2, more current data may produce different estimates for this age group. A determination will be made when we have a final set of county estimates based on the 2005-2007 CPS ASEC samples.

Based on current results, we recommend that the publication of estimates of low-income women aged 50-64 for counties not be actively pursued at this time. We will continue to research improvements in the model for possible future publication.

[8] As discussed in Section 1, this report addresses all states using the ≤ 200% IPR; if the model is successful at the 200% IPR, we expect that the estimates will be more reliable at the 250% IPR for states needing this information.